Spam Fighting in Social Tagging Systems

Sasan Yazdani[1], Ivan Ivanov[2], Morteza AnaLoui[1],
Reza Berangi[1], and Touradj Ebrahimi[2]

[1] Iran University of Science and Technology, Tehran, Iran
[2] École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
sasan_yazdani@comp.iust.ac.ir, {ivan.ivanov,touradj.ebrahimi}@epfl.ch,
{analoui,rberangi}@iust.ac.ir

Abstract. Tagging in online social networks is very popular these days, as it facilitates search and retrieval of diverse resources available online. However, noisy and spam annotations often make it difficult to perform an efficient search. Users may make mistakes in tagging and irrelevant tags and resources may be maliciously added for advertisement or self-promotion. Since filtering spam annotations and spammers is time-consuming if it is done manually, machine learning approaches can be employed to facilitate this process. In this paper, we propose and analyze a set of distinct features based on user behavior in tagging and tags popularity to distinguish between legitimate users and spammers. The effectiveness of the proposed features is demonstrated through a set of experiments on a dataset of social bookmarks.

Keywords: Social tagging systems, Social spam, Spam detection, Spammers, User behavior, Tags popularity.

1 Introduction

Social systems (networks) allow users to store, share, search and consume content (resources) online. Tagging in social systems has become increasingly popular since the transition to Web 2.0, as it simplifies and eases search and retrieval of information, and allows users to access these information globally while interact and collaborate with each other. Tags can be assigned to different types of resources, such as images, videos, publications and bookmarks, making it a valuable asset to search engines on the Internet and in social tagging systems.

A few challenges have been identified in research community as important in social tagging systems, namely tag recommendation, tag propagation and tag relevance. For example, *tag recommendation* approaches suggest appropriate tags to resources (e.g., videos) in order to make it easy for users to search and access information in social systems [11]. In order to speed up the time-consuming manual tagging process, tags can be automatically assigned to images by making use of *tag propagation* techniques based on the similarity between image content (e.g., famous landmarks) and its context (e.g., associated geotags) [7]. Since user-contributed tags are known to be uncontrolled, ambiguous and personalized, one of the fundamental issues in tagging is how to reliably determine

K. Aberer et al. (Eds.): SocInfo 2012, LNCS 7710, pp. 448–461, 2012.

the relevance of a tag with respect to the content it is describing [1]. The fact that tags are user-contributed enables spammers to pollute social systems with irrelevant or wrong information (spam) to mislead other users, and to damage the integrity and reliability of social systems. In general, spam on the Internet is created to trick search engines by giving the spam content higher rank in the search results for advertisement or self-promotion purposes. Various techniques have been proposed in the literature for combatting spam, for example, Google's PageRank [10] and TrustRank [20].

Tags play a vital role in social systems, since it is important that resources in these systems are assigned with relevant tags. Injection of irrelevant tags and inappropriate content in social systems can be performed mainly in two ways. First, spammers can use legitimate resources and assign irrelevant tags to them for the purpose of advertisement or self-promotion [3]. Second, spammers can use popular and high ranking tags to describe a spam resource and boost its rank [12]. Therefore, one of the most important issues in social tagging systems is to identify appropriate tags and at the same time filter or eliminate spam content or spammers.

In this paper, we propose a set of distinct features that can efficiently identify spam users in social tagging systems. The introduced features address various properties of social spam and users activities in the system, and provide a helpful signal to discriminate legitimate users from spammers. The effectiveness of the proposed features is demonstrated through a set of experiments on a dataset of social bookmarks.

The rest of the paper is organized as follows. Section 2 reviews the most recent related work. In Section 3, we propose a set of distinct features for spammer detection based on user behavior in tagging and tags popularity. Evaluation methodology and dataset are presented in Section 4. In Section 5, we compare several supervised learning approaches applied to the proposed features and analyze their performance. Finally, Section 6 concludes the paper with a summary and some perspectives for future work.

2 Related Work

The research work presented in this paper is related to different fields including tagging, tags characteristics, impact of spam and fighting spam in social systems. Therefore, the goal of this section is to review the most relevant work in the fields of spam impact on tagging and fighting against spammers in social tagging systems.

2.1 Tag Characteristics and Spam Impact in Tagging

Xu *et al.* [19] studied the characteristics of tags and categorized them into five groups: content-based which are used to describe the category an object belongs to, context-based which provide contextual information about the resource, attribute tags which point unnoticeable characteristic of a resource, subjective tags

which describe users point of view, and organizational tags that are personal like reminders and scheduler tags. Furthermore, the authors introduced criteria that must be fulfilled in order for a tag to be considered good. According to their study, a well-defined tag has properties like coverage of multiple facets of the resource, employing popular tags, excluding unlikely tags such as organizational or subjective tags.

Koutrika *et al.* [12] were the first to explicitly discuss methods of tackling spamming activities in social tagging systems. The authors studied the impact of spamming through a framework for modeling social tagging systems and user tagging behavior. They proposed a method for ranking content matching a tag based on taggers reliability in social bookmarking service Delicious. Their coincidence-based model for query-by-tag search estimates the level of agreement among different users in the system for a given tag. A bookmark is ranked high if it is tagged correctly by many reliable users. A user is more reliable if his/her tags more often coincide with other users tags. The authors performed a variety of evaluations of their trust model on controlled (simulated) dataset by populating a tagging system with different user tagging behavior models, including a good user, bad user, targeted attack model and several other models. Using controlled data, interesting scenarios that are not covered by real-world data could be explored. It was shown that spam in tag search results using the coincidence-based model is ranked lower than in results generated by, e.g. a traditional occurrence-based model, where content is ranked based on the number of posts that associate the content to the query tag.

2.2 Spam Fighting in Social Tagging Systems

Heyman *et al.* [5] classified anti-spam (or spam fighting) approaches into three categories: prevention-, rank- and identification-based. *Prevention-based* approaches employ series of mechanisms to keep spammers out of social tagging systems, such as CAPTCHA [16] and reCAPTCHA [17], or make it hard for spammers to pollute social system by restricting access, limiting number of resources a user can interact with, or requiring registration fee. Usually, prevention-based approaches are used as complementary defense systems to rank- or identification-based approaches. *Rank-based* approaches are very common in search by query scenarios and are used to demote spam content in order to return most legitimate resources on top of search results. *Identification-based* (or detection-based) approaches create a model from users' information, activities and interactions to efficiently detect and filter spam users (or content) from social tagging systems.

Bogers *et al.* [3] proposed an approach to identify spammers in social bookmarking systems such as BibSonomy and CiteULike. The approach is based on user language models assuming that spammers and legitimate users use different language jargons when posting. To detect spam users, they learned a language model for each post, and then measured its similarity to the incoming posts by making use of Kullback-Leiber (KL) divergence. The spam status of a new post takes the status of the most similar language model. Status of a user is determined by grouping all

users posts. This approach was evaluated on BibSonomy dataset for spam detection, proposed at ECML PKDD Discovery Challenge 2008 [6].

Krause *et al.* [13] employed a machine learning approach to detect spammers in BibSonomy. They investigated a framework for detecting spammers. The authors assumed that spammers usually use different strategies for polluting social bookmarking systems such as creating several accounts, publishing a particular post several times, and using semantically diverse tags to describe a bookmark and teaming up with other spammers to give good votes to each other. The authors investigated features considering information about a users profile, location, bookmarking activity and semantics of tags. By making use of these features, and naïve Bayes, support vector machine (SVM) classifiers, logistic regression and J48 decision trees, they were able to distinguish legitimate users from malicious ones. This study represents a good foundation for future machine learning spam detecting approaches.

Markines *et al.* [14] proposed six different tag-, content- and user-based features for automatic detection of spammers in BibSonomy. The authors used features representing the probability of a tag being spam, number of advertises per post and number of valid resources per user posts. It was shown that "TagSpam" feature (tag diversity in posts) is the best predictor of spammers among all other features, because spammers tend to use certain "suspect" tags more than legitimate users. Although their work showed promising results, most of the proposed features rely on an infrastructure to enable access to the content, and must be recalculated periodically to remain reliable. Therefore, the feasibility of the proposed features depends on the circumstances of a particular social tagging system.

Although BibSonomy is the most popularly explored domain for spam fighting, there are researchers who developed techniques for other social systems, such as Delicious, YouTube, MySpace or Twitter. Ivanov *et al.* [8] surveyed recent advances in techniques for combatting noise and spam in social tagging systems, classified the state-of-the-art approaches into a few categories and qualitatively compared and contrasted them.

3 Distinct Features

In this section, we first present a model of a social tagging system and then introduce a set of distinct features to distinguish between legitimate users and spammers in social systems.

Social tagging systems allow users to assign tags to resources shared online in order to enrich a resource with metadata and facilitate search for a particular resource, as previously explained in this paper. General model of a social tagging system is represented as a hyper-graph structure called *folksonomy* where the set of nodes consists of three kinds of objects: users, resources and tags, and hyper-edges connect these objects based on their relations [2]. The folksonomy can be defined with a quaternary structure $F = (U, T, R, P)$, where U represents the set of users u in the system, T is the set of tags t posted by users, R shows

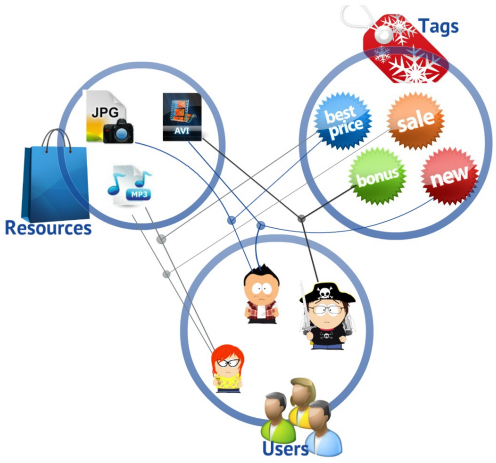


Fig. 1. An example of folksonomy representing a social tagging system with 3 users, 4 tags, 3 resources and 5 posts

the set of resources r and P defines the relation existing between tags, users, and resources. A relation linking a user, a tag and a resource represents a post. A post p in folksonomy can be represented with a triple $p = (u, r, T_u)$ which relates a user u who associated a resource r with a set of n tags $T_u = \{t_1, t_2, ..., t_n\}$. Figure 1 shows an example of folksonomy with 3 users, 4 tags and 3 resources.

Distinguishing between legitimate users and spammers in social tagging systems can be regarded as a classification problem. The most important part in any classification problem is the extraction of a good set of features from data. Features should represent data well to achieve good classification rate. Features are used to reduce the dimensionality of data while keeping important and relevant information. After studying the BibSonomy user behavior, we introduce 16 distinct features for each user from the evaluation dataset. Each user is represented with a feature vector consisting of 16 features which can be used by any known classifier to fight spam. In the following, we describe the proposed features in details, discuss the observation behind them and explain how to extract them out of a folksonomy.

3.1 LegitTags/SpamTags

We studied users behavior in BibSonomy and found out those spammers and legitimate users tend to use different languages for their posts. Spammers often use a fraction of legitimate user vocabulary, mostly popular tags, to gain higher ranks. Apart from this fact, they have a very distinctive jargon which is barely used by legitimate users. Based on these observations, we propose two features: *LegitTags* and *SpamTags*. *LegitTags* calculates the number of tags a user has posted which are mostly used by legitimate users. However, spammers also have habit to use popular tags that are previously posted by legitimate

users. Therefore, we introduce a feature *LegitTags* which defines the probability that a particular tag is used only by legitimate users. Let U_t be the set of all users in a social tagging system who associated at least one resource with a tag t, T_u be the set of all tags posted by a user u, S_t be a subset of spammers in U_t and L_t be a subset of legitimate users in U_t. Then, the feature *LegitTags* for user u can be calculated as follows:

$$LegitTags_u = \frac{1}{|T_u|} \sum\nolimits_{t \in T_u} \delta(u,t), \tag{1}$$

where $\delta(u,t)$ returns 1 if $|S_t|/|U_t|$ is less than a predefined threshold Th_{Legit}, otherwise it returns 0. Analogously, a feature *SpamTags* is defined as:

$$SpamTags_u = \frac{1}{|T_u|} \sum\nolimits_{t \in T_u} \sigma(u,t), \tag{2}$$

where $\sigma(u,t)$ returns 1 if $|L_t|/|U_t|$ is less than a predefined threshold Th_{Spam}, otherwise it returns 0. Optimal threshold values for Th_{Legit} and Th_{Spam} are experimentally found, and for our evaluation dataset they are set to 0.21 and 0.13, respectively.

3.2 Tags Popularity Based Features

One characteristic of spammers is that they tend to use popular tags when annotating online resources to gain higher rank in a search by keyword scenario [12], as already discussed in Sections 1 and 2. Based on this finding, we propose six features which address the popularity of tags shared in a social tagging system, namely, *LegitPopularity*, *SpamPopularity*, *TagPopularity*, *DistinctLegitPopularity*, *DistinctSpamPopularity* and *DistinctTagPopularity*.

For a particular tag t, we define a feature *LegitPopularity* as the number of times users in L_t used tag t in their posts. In an analogous way, features *SpamPopularity* and *TagPopularity* represent the number of times tag t was assigned to resources by users in S_t and U_t, respectively.

We propose three additional features representing tags popularity, namely *DistinctLegitPopularity*, *DistinctSpamPopularity* and *DistinctTagPopularity*. They represent the number of users in L_t, S_t and U_t who assigned tag t to at least one resource, respectively.

3.3 User Activity Based Features

User activity based features take advantage of user's posting behavior in a social system to better discriminate between legitimate users and spammers. These features are explained in the following and summarized in Table 1. All features are computed for each user separately.

Feature *AverageTagsPerPost* shows the average number of tags a user assigned to different resources. The rationale behind this feature is that posts from legitimate users usually have more tags describing resources compared to

Table 1. Summary of user activity based features. All features are computed for each user separately.

Distinct feature	Description
AverageTagsPerPost	Avg. no. of tags a user assigned to different resources
AverageDistinctTagsPerPost	Avg. no. of unique tags a user assigned to different resources
NewTags	No. of unprecedented tags a user added to the global dictionary of tags
Legit2Spam	Ratio between no. of legitimate and spam tags assigned by a user
TagsPerUser	Total no. of tags a user assigned to different resources
DistinctTagsPerUser	Total no. of unique tags a user assigned to different resources
Posts	No. of posts shared by a user
DistinctTagRatio	Ratio between no. of unique tags and total no. of tags assigned by a user

posts shared by spam users. With the same rational, we introduce a feature *TagsPerUser*, defined as the total number of tags a user assigned to different resources.

Based on our observation that spammers tend to use different popular tags for different posts and, at the same time, the intersection between sets of tags in two arbitrary posts from one spammer is none or very small, we introduce a feature called *AverageDistinctTagsPerPost*. This feature measures the average number of unique tags a user assigned to different resources. With the same rational, we present two other features: *DistinctTagsPerUser*, defined as the total number of unique tags a user assigned to different resources, and *DistinctTagRatio*, which represents the ratio between number of unique tags and total number of tags assigned by a user.

Furthermore, number of new tags introduced by spammers to the global dictionary of tags is relatively higher than number of tags introduced by legitimate users. Based on this fact, we introduce a feature *NewTags*. This feature is defined as the number of unprecedented tags a user added to the global dictionary of tags.

We present here two other user activity based features. A feature *Legit2Spam* represents the ratio between number of legitimate and spam tags assigned by a user, while a feature *Posts* is defined as the number of posts shared by a user.

Discussion on the performance of all proposed features on discriminating legitimate users from spammers is presented in Section 5.

4 Evaluation

In this section, we present a dataset and classification metrics used to evaluate the set of proposed features.

Table 2. Statistics of the original dataset (ECML PKDD Discovery Challenge 2008) and a reduced dataset used for evaluation

Statistics of datasets	Original dataset			Evaluation dataset		
	Legitimate	Spam	Total	Legitimate	Spam	Total
No. of users	2,467	29,248	31,715	500	500	1000
No. of resources	401,250	2,060,707	2,461,957	172,452	65,378	237,830
No. of tags	816,197	13,258,759	14,074,956	477,794	473,544	951,338
Avg. posts per user	162	70	77	344	131	238
Avg. tags per user	330	453	506	955	947	951
Avg. tags per post	2	7	6	3	8	4

4.1 Dataset

We used dataset collected from BibSonomy. BibSonomy is a social tagging system that allows users to share bookmarks and publication references. The system is aimed for researches and academic institutions which require a system without irrelevant information and commercial content. Therefore, this system has a rigorous policy against spammers. Moderators in this system manually find and remove spammers from the system [3]. If a user is labeled as a spammer, his/her posts will be no longer visible to other users. Spammer posts will not be removed from the system and this fact gives an illusion to spammers that they are still able to pollute the system.

We used a public dataset released by BibSonomy as a part of the ECML PKDD Discovery Challenge 2008 on Spam Detection in Social Bookmarking Systems [6]. Table 2 summarizes statistics of the dataset. This dataset consists of around 32,000 users who are manually labeled either as spammers or legitimate users, user-contributed tags and resources (bookmarks) which can be either web pages or BibTeX files. However, as shown in the second column of Table 2, an important skewness is present in this dataset since a majority of the users are spammers. This means that if a classifier labels all users as spammers, we would achieve a classification accuracy of over 0.92. Therefore, we selected randomly a subset of users (500 legitimate users and 500 spammers) to achieve a balance with respect to the number of users. Statistics of the dataset used for evaluation in this paper is shown in the third column of Table 2.

4.2 Classification Metrics

After having extracted proposed features from the evaluation dataset, several supervised classification methods, such as support vector machine (SVM), AdaBoost and decision trees, were applied on the extracted features to classify users as legitimate or spammers. Given the ground truth and the predicted labels, a confusion matrix is created and the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are computed.

Different metrics are used to evaluate the proposed features. The accuracy of the classification when shown solely is not a good indicator of a classifier behavior, and therefore, we calculated some complementary measures to thoroughly evaluate the proposed features. In addition to the classification accuracy defined as $\frac{TP+TN}{TP+FP+FN+TN}$, we calculated: (1) false positive rate (FPR) as $\frac{FP}{FP+TN}$, (2) precision (P) as $\frac{TP}{TP+FP}$, (3) recall (R) as $\frac{TP}{TP+FN}$, (4) F-measure as $\frac{2 \cdot P \cdot R}{P+R}$, and (5) area under receiver operating characteristics (AUC ROC) which represents the probability that an arbitrary legitimate user is ranked higher than an arbitrary spammer. Finally, we determined Matthews Correlation Coefficient (MCC) [15] to validate our result. As a less known performance metric, we explain it here in more details. MCC is a performance quality measure used in two-class classification problems. It is often used as a performance metric in bioinformatics. MCC is defined as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{3}$$

MCC has values between -1 and +1, where +1 indicates perfect classification (prediction), -1 shows total disagreement between prediction and observation, and 0 represents a random classification.

5 Discussion

In this section, we discuss the prominence of the proposed features for detection of spammers. First, performance of each feature separately is estimated and then some of them are aggregated to improve the classification performance. Finally, performance of different classifiers are compared and analyzed. All performance criteria were evaluated by making use of classifiers in Weka [18], a software library of most distinguished machine learning algorithms. Evaluation is performed using 10-fold cross-validation and default values for all parameters in Weka.

Figure 2 shows how well each of the proposed 16 features discriminates spammers. A decision stump classifier in Weka is applied on extracted features and the performance of each proposed feature is measured as accuracy, AUC ROC and F-measure. As we can see from the accuracy metric, each feature is able to correctly classify at least 60 % of users. Feature *LegitTags* has the best performance with more than 0.91 of accuracy in classification, and it is followed by *SpamTags*, *DistinctLegitPopularity* and *Legit2Spam* with 0.87, 0.76, 0.73 of accuracy, respectively. For classification of randomly selected users, as it can be seen from AUC ROC, again *LegitTags* and *SpamTags* have the best performance with 0.96 and 0.93 of AUC ROC. F-measure follows the trend of accuracy and AUC ROC, showing that *LegitTags* and *SpamTags* are the adequate features. Having considered all these measures, we can conclude that after *LegitTags* and *SpamTags*, tags popularity based features are the best performing set of features.

Feature *LegitTags* has the ability to very well separate spammers from legitimate users when fed solely into the classifier, as discussed previously in this

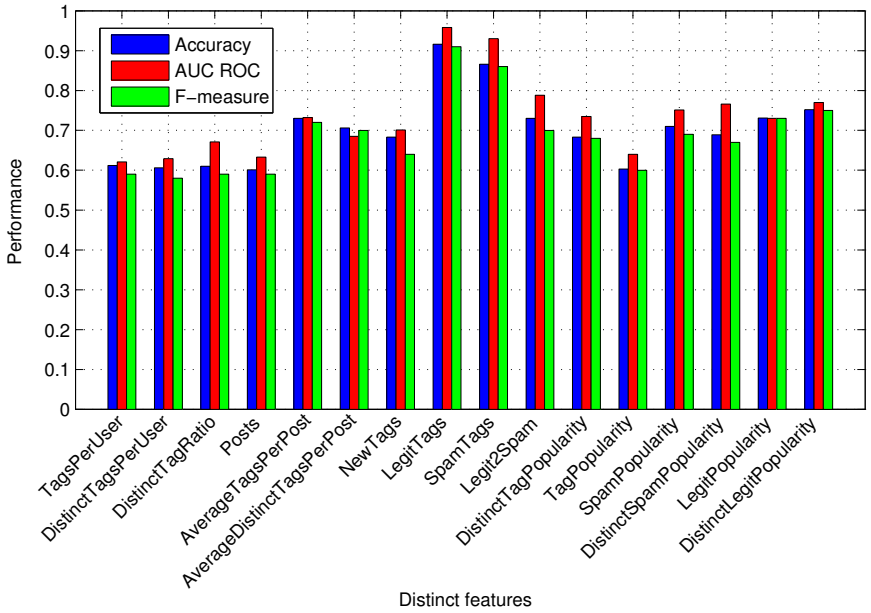
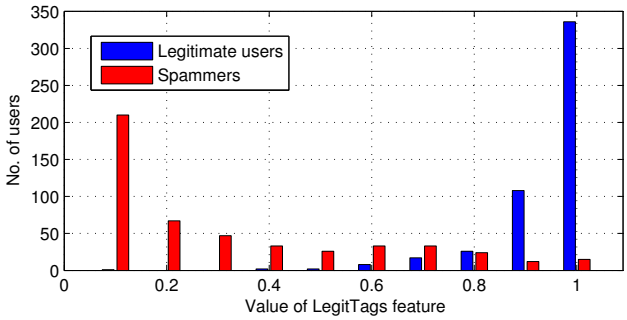


Fig. 2. The performance of each proposed feature plotted as accuracy, AUC ROC and F-measure

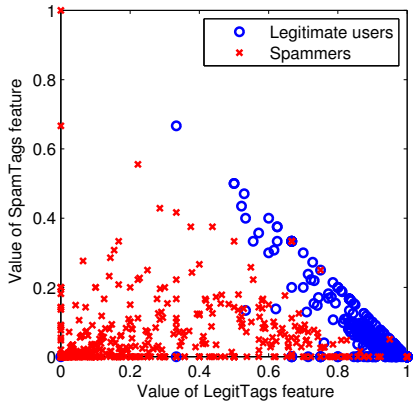
section. Therefore, we explore the performance of this feature in more details. 10-bins histogram of *LegitTags* values calculated from the evaluation dataset is shown in Figure 3 (a). When this feature is combined with the second best performing feature *SpamTags* and feature values are plotted in the feature space, we obtain the distribution shown in Figure 3 (b). These distributions give a visual intuition for how well feature *LegitTags* alone or combined with other feature separates two types of users. We can clearly see that the distributions of legitimate users and spammers can be easily separated by a simple threshold, for case (a), or line, for case (b). Therefore, linear discrimination classifiers are enough for spammers detection when using *LegitTags* and *SpamTags* features.

After *LegitTags* and *SpamTags*, tags popularity based features are the most powerful set of features, as shown in Figure 2. To further evaluate these features, we applied a standard discrimination function, the χ^2 statistics. The χ^2 (chi-square) statistics measures the goodness and powerfulness of features used for classification [9]. Again, we used Weka to apply this discrimination function. Figure 4 shows the consistent ranking of our six tags popularity based features to discriminate spammers from legitimate users.

It is well known that classification accuracy can be significantly improved by aggregating weak features rather than feeding different features separately into a classifier [4]. We can see from Figure 2 that each tag popularity and user activity based features have less than 0.75 and 0.73 of accuracy, respectively. Nevertheless, combination of these features results in a performance improvement.



(a)



(b)

Fig. 3. Discrimination power of the feature *LegitTags* to separate two types of users, when: (a) used alone, (b) combined with the feature *SpamTags*. Figure (a) represents the histogram of *LegitTags* values, and Figure (b) shows projection of *LegitTags* and *SpamTags* values in the feature space attempting to separate legitimate users (blue circles) from spammers (red crosses).

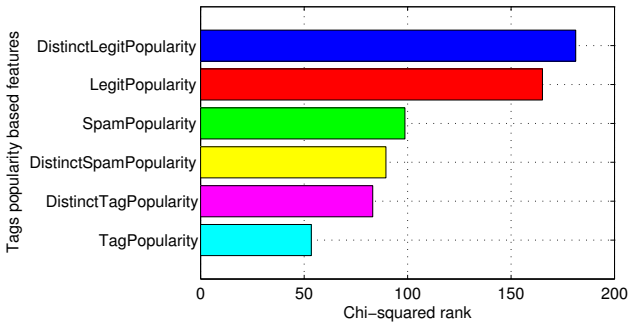


Fig. 4. Chi-squared ranking for all tags popularity based features

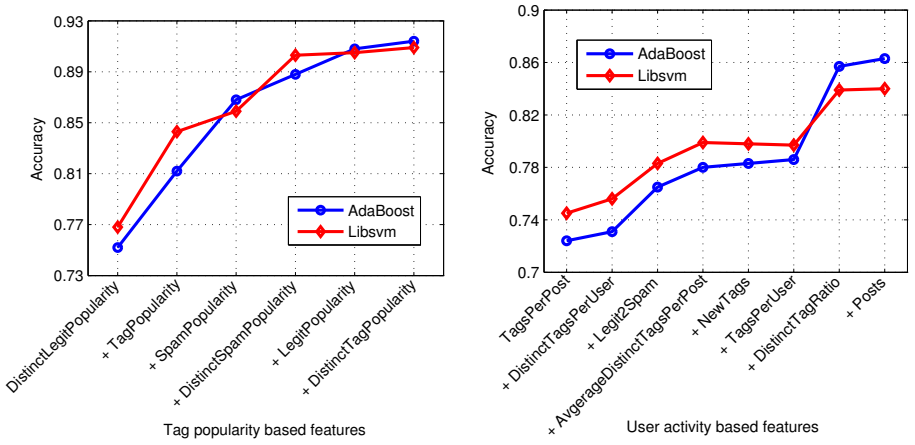


Fig. 5. Enhancement in the classification performance by aggregating: (a) all tag popularity based features, and (b) all user activity based features

Figures 5 (a) and (b) show how classification performance can be improved by separately aggregating tag popularity based features and user activity based features. Results are shown for two classifiers, namely AdaBoost and LibSVM. By combining all tag popularity based features we can improve classification accuracy from 0.75 to 0.91, while aggregating all user activity based features the accuracy increases from initial 0.73 to 0.86.

Finally, the proposed features are fed into more than 40 different classifiers and their performance in classification is evaluated. We used Weka to train classifiers with our features and to measure performance. Diverse classifiers are used, such as decision trees, neural networks and LibSVM, in order to have different perspectives on discriminative functions in feature space. Furthermore, ensemble classifiers [4] such as AdaBoost, bagging and rotation forest, were employed to have a comprehensive evaluation. The top 10 performing classifiers are reported in Table 3. Results show that AdaBoost was the best classifier for the evaluation dataset. It performs well with 0.987 of accuracy and only 0.013 of FPR. LibSVM and rotation forest classifiers have slightly lower accuracy of 0.986 and 0.981, with 0.014 and 0.019 of FPR, respectively. As noted by Markines *et al.* [14], in a deployed social spam detection system it is more important that FPR is kept low compared to high accuracy, because misclassification of a legitimate user is a more consequential mistake than missing a spammer. Other researchers, who proposed different features from the whole or partial dataset of ECML PKDD Discovery Challenge 2008, obtained similar results, for example, Markines *et al.* [14] were able to reach 0.979 of accuracy and 0.013 of FPR, while Bogers *et al.* [3] got 0.9799 of classification accuracy.

Table 3. Top classifiers created in Weka. Evaluation is performed using 10-fold cross-validation. The best performing classifier and metric values are highlighted in **bold**.

Weka classifier	Accuracy	FPR	R	P	F-measure	AUC ROC	MCC
AdaboostM1	**0.987**	**0.013**	**0.994**	0.980	**0.987**	0.993	**0.974**
Libsvm	0.986	0.014	0.978	**0.994**	0.986	0.993	0.973
RotationForest	0.981	0.019	0.981	0.978	0.980	0.993	0.962
SMO	0.979	0.021	0.979	0.979	0.979	0.991	0.958
RBFNetwork	0.975	0.025	0.965	0.986	0.975	0.993	0.95
Bagging	0.974	0.026	0.974	0.974	0.974	**0.996**	0.948
Decorate	0.973	0.029	0.970	0.968	0.968	0.990	0.930
FT	0.972	0.028	0.966	0.972	0.970	0.985	0.944
MultiBoostAB	0.971	0.029	0.970	0.972	0.971	0.987	0.942
MLP	0.971	0.029	0.959	0.984	0.971	0.982	0.942

6 Conclusions

In this paper, we presented different features suitable for fighting spam in social tagging systems. The problem of having trustworthy tags associated to resources is important in social systems, because of their increasing popularity as means of sharing interests and information. Therefore, one of the most important issues in social tagging systems is to identify appropriate tags and at the same time filter or eliminate spam content or spammers.

We proposed 16 distinct features based on user activity in posting and tags popularity. The prominence of the proposed features in distinguishing between legitimate users and spammers is discussed. We measured the performance of each feature solely and showed that *LegitTags* feature, defined as the probability that a particular tag is used only by legitimate users, performed the best. We also showed that aggregation of features leads to the improvement in the classification performance. Finally, performance of different classifiers was compared. The results are promising. The best classifier achieved accuracy of 0.987 with false positive rate of 0.013 in discriminating legitimate users from spammers.

As a future study, we will explore more sophisticated features which are able to deal with dynamics of trust, by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling of phenomenon in real-world applications.

Acknowledgment. This work was supported by the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2).

References

1. Xirong, L., Snoek, C., Worring, M.: Learning tag relevance by neighbor voting for social image retrieval. In: Proc. ACM MIR, pp. 180–187 (2008)
2. Benz, D.K., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The social bookmark and publication management system BibSonomy. VLDB Journal 19(6), 849–875 (2010)
3. Bogers, T., Van den Bosch, A.: Using Language Models for Spam Detection in Social Bookmarking. In: Proc. ECML/PKDD Discovery Challenge, pp. 1–12 (2008)
4. Duda, R., Hart, P.: Pattern classification and scene analysis. Wiley (1973)
5. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing 11(6), 36–45 (2007)
6. Hotho, A., Benz, D., Jäschke, R., Krause, B.: ECML PKDD Discovery Challenge (2008), `http://www.kde.cs.uni-kassel.de/ws/rsdc08`
7. Ivanov, I., Vajda, P., Jong-Seok, L., Goldmann, L., Ebrahimi, T.: Geotag propagation in social networks based on user trust model. MTAP 56(1), 155–177 (2012)
8. Ivanov, I., Vajda, P., Jong-Seok, L., Ebrahimi, T.: In tags we trust: Trust modeling in social tagging of multimedia content. IEEE SPM 29(2), 98–107 (2012)
9. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Proc. ICTAI, pp. 338–391 (1995)
10. Rogers, I.: The Google PageRank algorithm and how it works (2002)
11. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag Recommendations in Folksonomies. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 506–514. Springer, Heidelberg (2007)
12. Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems. In: Proc. ACM AIRWeb, pp. 57–64 (2007)
13. Krause, B., Schmitz, C., Hotho, A., Stumme, G.: The anti-social tagger: Detecting spam in social bookmarking systems. In: Proc. ACM AIRWeb, pp. 61–68 (2008)
14. Markines, B., Cattuto, C., Menczer, F.: Social spam detection. In: Proc. ACM AIRWeb, pp. 41–48 (2009)
15. Matthews, B.W.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta 405(2), 442–451 (1975)
16. Von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: Using Hard AI Problems for Security. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 294–311. Springer, Heidelberg (2003)
17. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-based character recognition via web security measures. Science 321(5895), 1465–1468 (2008)
18. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005), `http://www.cs.waikato.ac.nz/ml/weka`
19. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. In: Proc. ACM WWW, pp. 1–8 (2006)
20. Gyongyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with TrustRank. In: Proc. VLDB, pp. 576–587 (2004)